

Ian Gent and Catherine Jones

- Interviewer Hello, and welcome to this Jisc podcast interview. In this series we're speaking to people working on projects being funded by Jisc's research data spring project to find out more about what they're developing. In this podcast I'm chatting with Catherine and Ian who are working on a project to work on the re-use, re-purposing and re-produceability of software.
- Ian My name's Ian Gent and I'm at the University of St Andrews in Scotland. I'm in the department of computer science, so it's quite a new experience in the world of data and Jisc kind of world.
- Catherine Hello, I'm Catherine Jones and I work at the Science and Technology Facilities Council in the scientific computing department, and I lead a group who do software engineering.
- Interviewer Thank you. Could you tell us what your project's called and explain what it's going to do?
- Catherine Yes, our project's called Software Re-use, Re-purposing and Reproduceability and it's looking at how software that supports research can be re-used by other people and found, and in the longer term remain usable for others to see how it's used.
- Interviewer Who else has been working with you on it?
- Catherine We've had some support from Brian Matthews who also works at STFC, and some colleagues at St Andrews.
- Ian Yes, so there's a colleague of mine called Simon Dobson, and also we eventually had a couple of really, really good interns, one in St Andrews, one at STFC: Charles Roux has just finished his degree and he's been doing some excellent stuff, and Paulina Latch, and we had Paulina up to visit St Andrews, so that's worked very well.
- Interviewer Why did you choose to tackle the particular problem you're looking to address?
- Catherine I think it's important in the data arena to remember that data just doesn't arrive and most data depends on specialist software to be understandable. So that's my particular interest in software and, as someone who originally was a software developer, I'm interested in ensuring it stays alive. Ian has a long term interest in recomputation.
- Ian Yes, so this is what some people would on a good day call a bench scientist. So one of the things that I'm very interested in is... So I don't come from this world, as I said, of data science and this kind of world that Catherine and most other people involved in Jisc come from, but my interest is in, I've run an experiment on my computer and I want somebody else to be running it again to check my results, and then that led to a website called recomputation.org which is focused on that, and then that led to my interest in

the Jisc Data Spring and we met Catherine. The particular focus of this project is to extend that a little bit, because as well as one particular experiment I'm very interested in the notion that you have maybe a lab working on, doesn't have to be computing, it could be any discipline, and it doesn't have to be science, and they're doing some analysis and they've got a computer programme that does the analysis, and then the PhD student who wrote it leaves and how do you carry on using that software, which maybe is absolutely brilliant but only works when that PhD student's at the keyboard? So that's one of the things we're aiming at.

Interviewer Catherine, have you had similar experiences?

Catherine Here at STFC we support people with their data but, as someone who is a computing professional, I'm aware that home grown software only tends to have a life of say three to five years and then you have to move on, and if you're trying to keep data for longer then there are issues around keeping the software alive, and if you need the software to interpret the data, or you've similar data and you want to use the software to interpret it and you don't want to write your own software, then it's very hard to find academic software that someone else has written. Because part of this project is about being able to find software that's already written.

Ian Yes, that's a really good point Catherine made, because the discovering side of it is really important and that's one I wasn't focusing on previously, so I'm really pleased to be looking at that as well.

Interviewer Could you tell me a little bit more about the work that you're doing day-to-day?

Ian One of the things, it sounds incredibly simplistic but one of the things I like to think about is like YouTube where, for software basically, so with YouTube somebody uploads a video to YouTube and then what you get in the future, if nobody deletes it, is that anyone can find that video by typing in cat, laughing, whatever it is they want to find, and then they go to that page and then there's just a button which says play on it, and they can play the video. And it's ridiculously overambitious and we're only going to do a small part of this, but we'd like a similar kind of thing, not necessarily YouTube, but somebody has some scientific software for [inaudible 4:58] analysis or something, or somebody doesn't have it, they want to find it, so they go to some search engine, they find it, they find a web page which has been built or provided with the tools we're talking about, and then maybe even on that, ideally on that web page, there's a button which says play and they can even start using it, maybe as a demo or something like that. Because obviously, if it's going to be a huge amount of time, they have to provide their own computing.

Catherine So Ian's expressed the vision, what we've done practically to date is start off with for this to work you have to be able to identify your software. So we've done some guidelines about assigning persistent identifiers to software, what is the thing you're trying to capture, and then the students, Charles and Paulina, have been working on techniques using standard things like Vagrant and Docker to encapsulate the software in a state that it's runnable. We're not writing software to do stuff in a big way, we're just showing that this way of encapsulating software so it can have a play button works once you've identified what it is exactly you're trying to encapsulate, because persistently identifying it means you have to assign meta data to it that will enable people to find it. So, for example, you might choose to use data site DOIs to persistently

identify your software, therefore you are enabling it to be discoverable through dataverse.

- Interviewer It sounds like you're casting your net really widely to find all kinds of different software; has anything cropped up or have any needs cropped up that have surprised you?
- Ian So it's not exactly the needs but you're coming from these different angles, so whenever I pontificate, like I did a couple of moments ago, and Catherine says, well yes, remember people have got to be able to discover it and the data citation principles and so on, so that's really good because otherwise you just go down this one avenue. I can have a bit of a single track mind, so it's really good that somebody else says, but what about that other side of it? So it's not exactly the needs of it, but it is the needs in the sense that people have to be able to discover it, the persistent identifier, which I didn't mention, which Catherine did, that's really good because, again, and that's not something you get with YouTube but it's like once a link to some piece of software exists, hopefully you should be able to use that same link, not a link, persistent identifier, hopefully forever.
- Catherine We've been exploring how to apply the technology within our own institutions; what we're hoping to do in this coming phase is explore this with another institution, and I think in that case we might get some insights that we don't have yet, because at the moment it's very project team focused.
- Interviewer I was going to ask you about the stage that you're at in the project now; how have you found the process of being part of the Research Data Spring to date?
- Catherine I think it's great, because I wouldn't have met Ian otherwise.
- Ian I think it's terrible because I wouldn't... No wait, sorry. I think it's great because I wouldn't have met Catherine otherwise, I completely agree, yes.
- Interviewer Has it given you opportunities to do things that you genuinely haven't done before, or is it just those new connections that create the interest?
- Catherine I think, as well as creating new connections, it's given me the opportunity to actually do something in an area I'm interested in that is of wider interest to the academic community.
- Ian Yes, and from my point of view, again the symmetric things as more a member of the academic community directly day-to-day, it's good to interact with people who can, like Catherine and other people found through Jisc, who can take the ideas to other places.
- Interviewer Is that what you've enjoyed most, or is it the practicalities of the project you've been enjoying getting your hands into?
- Ian I think, well for me, definitely meeting up with Catherine and Brian and people like that; well not really people like that, them! But also we met other people at the first workshop. I enjoyed the bouncing ideas around at that workshop, and the work we've been doing, I mean it's wonderful what the interns have been doing, but seeing what they're doing, which some of that to me is amazing, and they're just sort of gluing tools together that exist and then doing things which seem amazing to me. That's really

good, and you see it's like, and you're blown away, which is obviously [inaudible 09:41] thing about them being good but you ask them to do interesting things and then it turns out to be interesting, so that's really nice.

Catherine I've enjoyed being part of a wider programme, meeting people and then going away and all doing stuff, and coming back and hearing what other people have done, so that's been good. As I say, part of a bigger research community.

Interviewer Can you tell us a little bit about what you're doing next?

Catherine So two aspects: with the guidelines about persistent identification we're involved with the [Force 11 10:12] software Citation group, so we're hoping to contribute to that and also get some feedback in the wider academic community about whether our suggestions make good sense, and then with the tools we're hoping to build on the work that the interns have done with a case study.

Interviewer How are you hoping your project will benefit people?

Catherine Well hopefully the guidelines about persistent identification will help people to think about, when they write software, that supporting the research cycle, and there is a research output in its own right, think about how it remains long-lived, because if you're assigning a long term identifier to it you have to think about things, and hopefully that will make an impact on people who write software.

Ian Exactly, and I particularly look at academics and especially, I think if it's a large group or a large community worldwide in a particular specialist area of software, they're probably pretty good at maintaining their code base, but I think I'm really interested in trying to help the ones where it's a small group, maybe a professor or an academic and PhD student, post-doc, and somebody moves on and you've got this almost... There's a lot of research groups which work by almost apostolic succession, which is one PhD student writes some code, the next one edits it, the next one, and you get to this point where if they ever have a point where there's no PhD student, that research area would die, just because of the software maintenance. So I really want to be able to help people in these much smaller groups, individuals or small groups have their software carry on working, and then it will help them with their research, and also hopefully help other scientists build on it, which at the minute is very difficult, because if I've got my software and my PhD student doesn't go to work with you, you're not going to be able to run my software anymore and build on it.

Catherine I think it also helps support the notion of software a research output in its own right, and citing software and getting credit for writing software, so assigning a persistent identifier is the first stage in that, because you have to know what you're citing.

Ian I completely agree, that's really important, and hopefully it would help, I think it's part of a much bigger effort of people, there's a phrase come into recent usage called research software engineer, so people may be in an academic environment where their job isn't to publish papers but to write software, and to get them proper credit for that is really important. This is just a small part of that, but it's an important effort.

Interviewer You've been working with the project since its inception, it's been a little while now you've been working together, could you describe the project in three words, each?

- Ian I was going to go with re-use, re-purposing and re-usability, but that might be cheating.
- Catherine I think it has been interesting, innovative and rewarding.
- Interviewer It sounds like you've got quite a large team going with various researchers and PhD students and lots of bits of kit and software that you've been finding out about, have any nicknames sprung up as you've been working together, for things or people?
- Catherine Well I was going to say calling it the three Rs is about... It didn't really have a name to start with it, did it?
- Ian Yes, and we haven't actually been... We have got a site called RR, with RR on it, but yes, you've been using the name three Rs, and yes, we maybe need to build on that as a brand.
- Interviewer If your project was an animal, what would it be?
- Ian I've kind of got an answer for this; I was thinking about this, but I'm not actually sure if it's an animal, because it's a thing you see in American usage, but I don't know if it's an actual animal, which is a pack rat; it's still not quite right, because a pack rat is something which, is somebody in American usage at least who kind of just keeps everything, and I want to keep all the software running. Catherine will need to be something else who can find the stuff that the pack rat has put away.
- Catherine A meerkat, standing up and surveying the surrounding while looking after its burrow. I'm not very good at zoologically, so that's...
- Ian That's very good, I like that, yes.
- Interviewer Thanks to Catherine Jones and Ian Gent for talking to us about their Three Rs for software, one of the projects being funded by Jisc's Research Data Spring project. If you'd like to find out more about it or any of the other projects taking part all the information is available online, just go to Jisc's website and search for research data spring to find the project pages.

[ENDS]